# A Survey On Document Summarization

Presenter: Chen Yi-Ting

# Reference(1/2)


- Lin-shan Lee and Berlin Chen, “Spoken Document Understanding and Organization,” IEEE Signal Processing Magazine (IEEE SPM), Vol. 22, No. 5, Sept. 2005
- T. Kikuchi, S. Furui, and C. Hori, “Two-stage automatic speech summarization by sentence extraction and compaction,” in Proc. IEEE and ISCA Workshop on Spontaneous Speech Processing and Recognition, 2003, pp.207-210.
- Sadaoki Furui, Tomonori Kikuchi, Yousuke Shinnaka, Chiori Hori, “Speech-to-Text and Speech-to-Speech Summarization of Spontaneous Speech”, IEEE transactions on speech and audio processing, VOL. 12 No.4, July 2004.
- Makoto Hirohata, Yousuke Shinnaka, Koji Iwano and Sadaoki Furui, “Sentence Extraction-Based Presentation Summarization Techniques and Evaluation Metrics”, ICASSP 2005.
- Y. Gong and X. Liu, “Generic text summarization using relevance measure and latent semantic analysis,” in *Proc. ACM SIGIR Conference on R&D in Information Retrieval*, 2001, pp. 19-25.
- 何遠，『中文口語文件自動摘要之初步研究』，碩士論文，國立臺灣大學電信工程學研究所，2003。
- 黃耀民，『以字句擷取爲基礎並應用於文件分類之自動摘要之研究』，碩士論文，國立臺灣師範大學資訊工程研究所，2005。
- J. Goldstein, M. Kantrowitz, V. Mittal and J Carbonell, “Summarizing text documents: sentence selection and evaluation metrics,” in *Proc. ACM SIGIR Conference on R&D in Information Retrieval*, 1999, pp. 121-128.

# Reference(2/2)


- M. Witbrock and V. Mittal, “Ultra-summarization: A statistical approach to generating highly condensed non-extractive summaries,” in *Proc. ACM SIGIR Conf. R&D in Information Retrieval*, 1999, pp. 315–316.
- Inoue A., Mikami T., Yamashita Y. *Improvement of Speech Summarization Using Prosodic Information* Proc. of Speech Prosody 2004, Japan.
- Konstantinos Koumpis, Steve Renals, “Automatic Summarization of Voicemail Messages Using Lexical and Prosodic Features”, ACM transactions on Speech and Language Processing, Vol. 2, No. 1, February 2005, Article 1.
- Chung-Hsien Wu, Chien-Lin Huang and Chia-Hsin Hsieh, “Spoken Document Summarization And Retrieval For Wireless Application”.
- Chien-Lin Huang, Chia-Hsin Hsieh and Chung-Hsien Wu, “Spoken Document Summarization Using Acoustic, Prosodic And Semantic Information”.
- Sameer Maskey, Julia Hirschberg, “Comparing Lexical, Acoustic/Prosodic, Structural and Discourse Features for Speech Summarization”, Interspeech 2005. 2006.
- Sameer Maskey,Julia Hirschberg, “Summarizing Speech without Text Using Hidden Markov Models”, HLT/NAACL 2006.
- Chin-Yew Lin and Eduard Hovy, “Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics”, *In Proceedings of the Human Technology Conference 2003 (HLT-NAACL-2003), May 27 . June 1, 2003, Edmonton, Canada.*
- C.Y. Lin, “ROUGE: Recall-oriented Understudy for Gisting Evaluation,” 2003, http://www.isi.edu/~cyl/ROUGE/.
- C.-Y. Lin, “Looking for a few good metrics: ROUGE and its evaluation,”Working Notes of NTCIR-4 (Vol. Supl. 2), pp. 1-8 (2004)

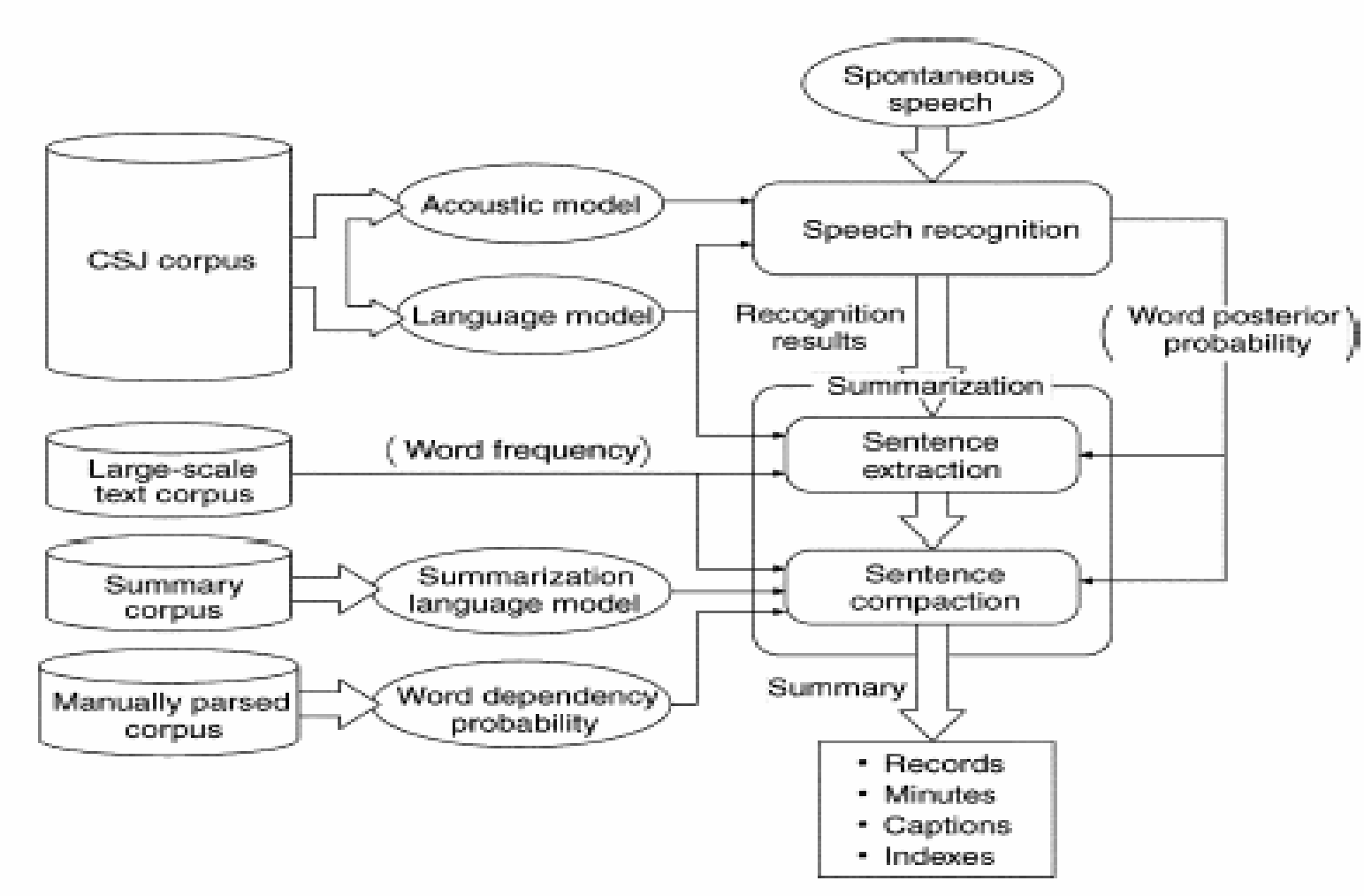
# Outline

- Introduction
- Overview of extractive summarization methods
- Evaluation Metrics
- Relevance Model
- Data Corpus and Experiments
- Conclusions

# Introduction

- The World Wide Web not only led to a renaissance in this area but extended it to cover a wider range of new tasks, including multidocument, multilingual, and multimedia summarization
- In general, the summarization can be either extractive or abstractive
- Text summarization Vs. Spoken document summarization
  - Speech summarization extracts important information and removes redundant and incorrect information form recognizing speech.
  - One fundamental problem with the summaries produced is that they contain recognition errors and disfluencies.
- Summarization presentation：Text presentation Vs. Speech presentation
- Single Document Summarization Vs. Multidocument summarization
- Summarization Evaluation：Subjective Evaluation Vs. Objective Evaluation

# Overview of extractive summarization methods(1/12)

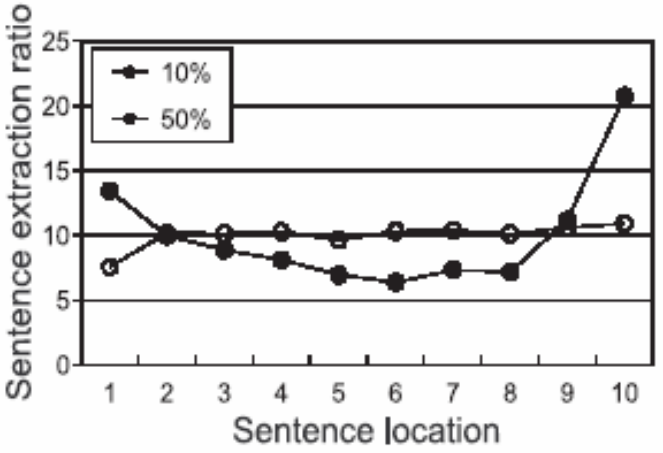


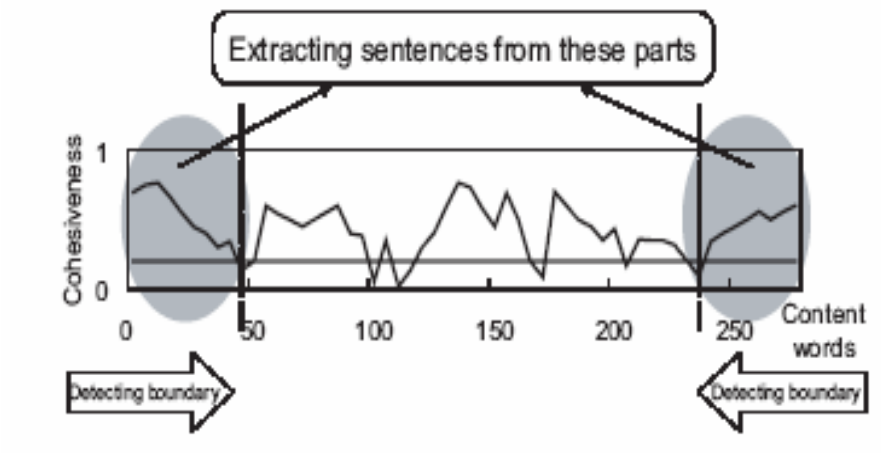
# Overview of extractive summarization methods(2/12)

- Extraction using sentence location/structural
- Extraction using degree of similarity (Cosine)
- Extraction using features score
- Extraction using latent semantic analysis
- Extraction using probabilistic model

# Overview of extractive summarization methods(3/12)

- Extraction using sentence location/structural
  - Sentence extraction using sentence location
  - Focusing on the introduction and conclusion segments
  - Specific structure on some domain
    - Ex: Broadcast News programs—sentence position、speaker type、previous-speaker type、next-speaker type、speaker change



# Overview of extractive summarization methods(4/12)

- Extraction using degree of similarity (Cosine)
  - Vector Space Model (VSM)
  - Relevance Measure (RM)
  - Indexing feature：word、character、syllable、word pair、character pair、syllable pair….
  - Weight：TF、TFIDF

$$TF : f_{i,d} = \frac{freq_{i,d}}{\max_{l} freq_{l,d}}$$

$$TFIDF : w_{i,d} = f_{i,d} \times \log \frac{N}{n_i}$$

$$sim(d_j, \vec{S}_k) = \frac{\vec{d_j} \cdot \vec{S}_k}{\left|\vec{d_j}\right| \times \left|\vec{S}_k\right|} = \frac{\sum_{i=1}^{L} w_{i,j} \times w_{i,s_k}}{\sqrt{\sum_{i=1}^{L} w^2_{i,j}} \times \sqrt{\sum_{i=1}^{L} w^2_{i,s_k}}}$$

# Overview of extractive summarization methods(5/12)

- Extraction using features score
  - Statistical measure (such as TF/IDF、significance score、significance score Using LSI)

$$I(w_i) = f_i \log \frac{F_A}{F_i} \quad , \quad R(w_i) = \max_b \left\{ P_{LSI}(w_i, w_b^{t^*}) \square f_{w_i} \square \ln(N / df_{w_i}) \right\}$$

  - Linguistic measure／Lexical Features (NEs and POSs)
    - Ex: total number of named entities in a sentence
  - Language model scores $L(w_i) = \log P(w_i \mid ....w_{i-1})$
  - Confidence scores
  - Prosodic/Acoustic Features
    - Ex: speaking rate、F0 minimum、F0 maximum、RMS energy、sentence duration

$$S(W) = \frac{1}{N} \sum_{i=1}^{N} \left\{ L(w_i) + \lambda_I I(w_i) + \lambda_C C(w_i) \right\}$$

# Overview of extractive summarization methods(6/12)

- Two-Stage Summarization Method (2004 SAP) :
  - Important sentence extraction
    - Score for each sentence $W = w_1, w_2, ....., w_N$

$$S(W) = \frac{1}{N}\sum_{i=1}^{N}\left\{L(w_i) + \lambda_I I(w_i) + \lambda_C C(w_i)\right\}$$

    - **Linguistic score**: $L(w_i) = \log P(w_i \mid ....w_{i-1})$
    - **significance score**: $I(w_i) = f_i \log \frac{F_A}{F_i}$
    - **confidence score**
  - Sentence Compaction
    - The remaining transcription is automatically modified into a written editorial article style.

# Overview of extractive summarization methods(7/12)

- Two-Stage Summarization Method (2004 SAP)：
  - Sentence Compaction

    - $L(w_i)$、$I(w_i)$、$C(w_i)$ and $T(w_i, w_j)$
    - $T(w_i, w_j)$, this score is a measure of the dependency between two words and is obtained by a phrase structure grammar, SDCFG
    - A set of words that maximizes a weighted sum of these scores is selected according to compression ratio and connected to create a summary using a two-stage DP technique

# Overview of extractive summarization methods(8/12)

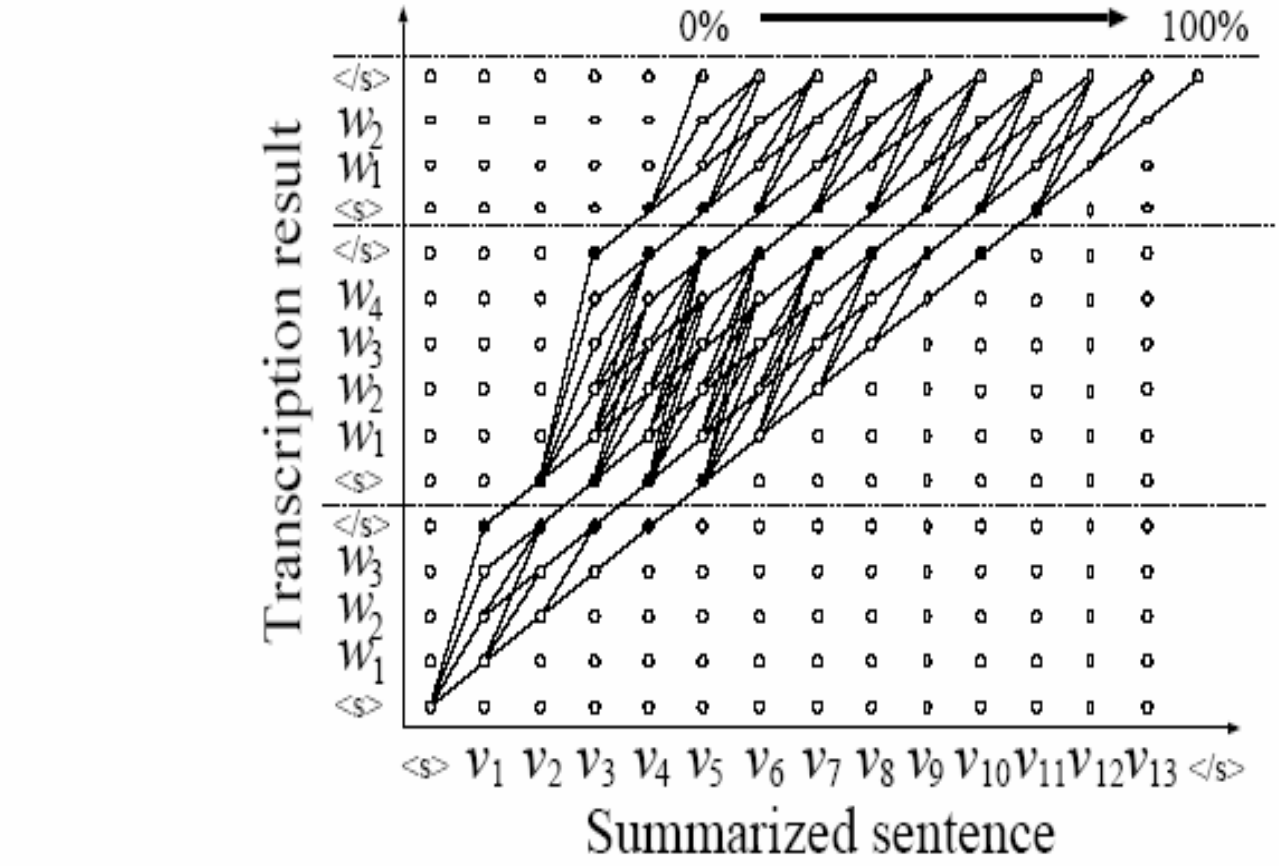
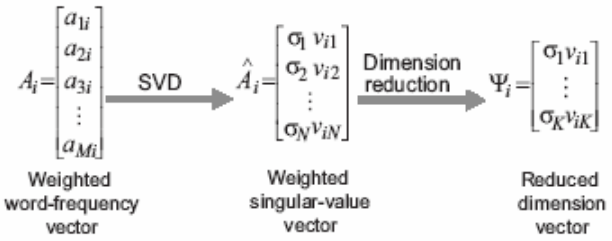


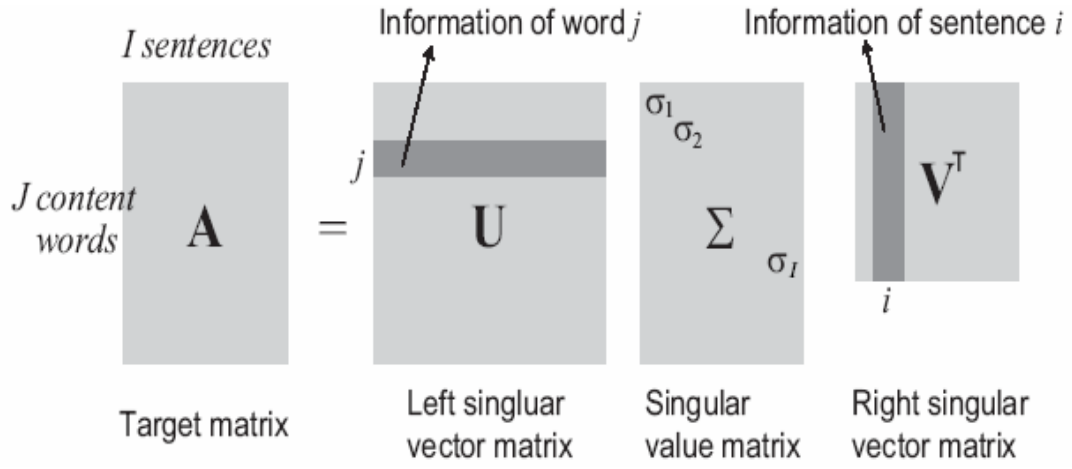

Figure 2.13: An example of DP process for summarization of multiple utterances.

# Overview of extractive summarization methods(9/12)

- Extraction using latent semantic analysis
  - Latent Semantic Analysis (LSA)
  - embedded Latent Semantic Analysis (eLSA)
  - Dimension reduction based on SVD (Furui LSA)

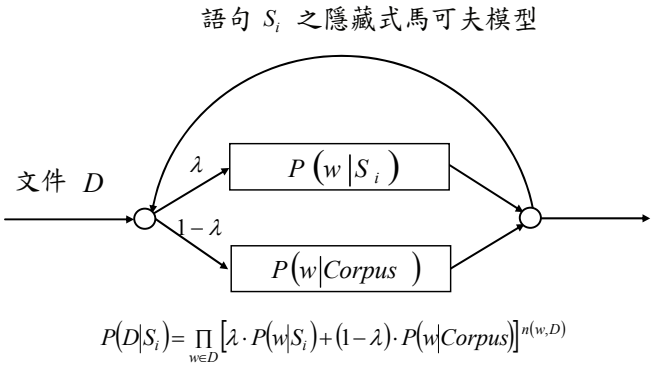
$$Score(i) = \sqrt{\sum_{k=1}^{K} (\sigma_k v_{ik})^2}$$



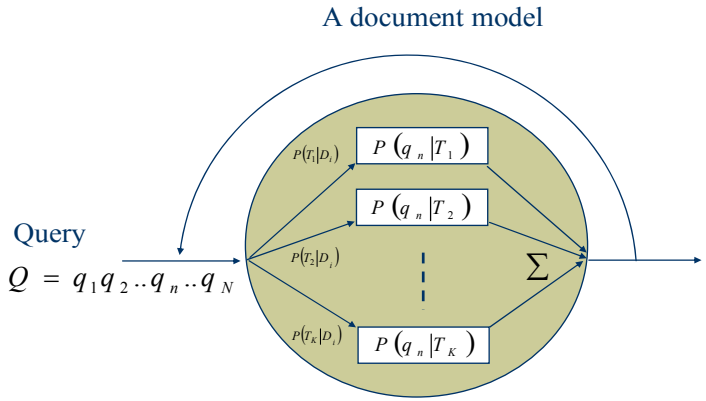
# Overview of extractive summarization methods(10/12)

- Extraction using probabilistic statistics
  - Hidden Markov Model (HMM):
    - Each sentence of a document was treated as a probabilistic generative model for predicting the document
    - which were directly estimated from each sentence itself and smoothed by *N* -gram distributions estimated from a large text corpus.
    - $P(D|S_i) = \prod_{w \in D} \left[\lambda \cdot P(w|S_i) + (1-\lambda) \cdot P(w|Corpus)\right]^{n(w,D)}$



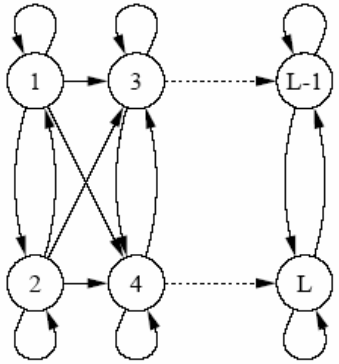
# Overview of extractive summarization methods(11/12)

- Extraction using probabilistic statistics
  - Topical Mixture Model (TMM )
    - Each sentence can be interpreted as a probabilistic generative topical mixture model (TMM).
    - In this model, a set of latent topical distributions characterized by unigram language models are used to predict the document terms, and each of the latent topics is associated with a sentence specific weight.
    - $$P\left(D\middle|S_i\right)=\prod_{w\in D}\left[\lambda\cdot P\left(w\middle|S_i\right)+\left(1-\lambda\right)\cdot\sum_{k=1}^{K}P\left(w\middle|T_k\right)P\left(T_k\middle|S_i\right)\right]^{n(w,D)}$$



# Overview of extractive summarization methods(12/12)

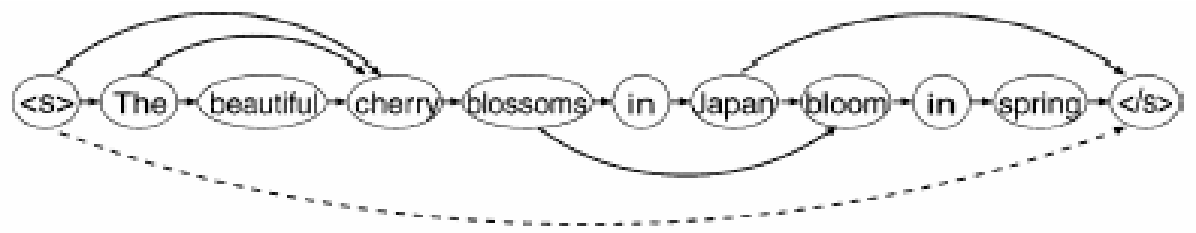
- Summarizing Speech Without Text Using HMM
  - Hidden Semi-Markov Model (HSMM)
  - Building position-sensitive HMM
  - L number of bins, 2L states
  - Features and Training




- Literal Term Matching Vs. Concept Matching
  - Literal Term Matching：
    - HMM、
    - Extraction using degree of similarity、
    - Extraction using features score
  - Concept Matching：
    - TMM、
    - Extraction using latent semantic analysis

# Evaluation Metrics(1/4)

- Objective Evaluation Metrics
  - **Summarization accuracy**
    - All the human summaries are merged into a single word network.
    - Word accuracy of the automatic summary is then measured as a summarization accuracy in comparison with the closest word string extracted from the word network.
    - Problem: the variation between manual summaries is so large. (high summarization ratio or low summarization)



# Evaluation Metrics(2/4)

- Objective Evaluation Metrics
  - **Sentence recall/precision**
    - Sentence recall/precision is commonly used in evaluating sentence-extraction-based text summarization.
    - Since sentence boundaries are not explicitly indicated in input speech, estimated boundaries based on recognition results do not always agree with those in manual summaries.
    - F-measure, F-measure/max, F-measure/ave.

$$R = \frac{|S_{man} \cap S_{sum}|}{|S_{man}|}, \quad \mathrm{P} = \frac{|S_{man} \cap S_{sum}|}{|S_{sum}|}, \quad \mathrm{F} = \frac{2\mathrm{RP}}{\mathrm{R}+\mathrm{P}}$$

# Evaluation Metrics(3/4)

- Objective Evaluation Metrics
  - **ROUGE-N**
    - ROUGE-N is an N-gram recall between an automatic summary and a set of manual summaries.
    - $$ROUGE-N = \frac{\sum_{S \in S_H} \sum_{g_n \in S} C_m(g_n)}{\sum_{S \in S_H} \sum_{g_n \in S} C(g_n)}$$
    - $S_H$ is a set of manual summaries, S is an individual manual summary, $g_n$ is an N-gram, $C(g_n)$ is the number of $g_n$'s in the manual summary, and $C_m(g_n)$ is the number of co-occurrences of $g_n$ in the manual summary and automatic summary.

# Evaluation Metrics(4/4)

- Objective Evaluation Metrics
  - **Cosine Measure**

$$ACC(m\%) = \frac{1}{|\{D\}|} \sum_{D \in \{D\}} \frac{\sum_{z=1}^{Z} SIM\left(A_D(m\%), E_{z,d}(m\%)\right) + SIM\left(A_D(m\%), R_{z,d}\right)}{6}$$

- Subjective Evaluation Metrics (Direct evaluation)
  - By human subjects
- Automatic evaluation
  - summaries are evaluated by IR

# Relevance Model

- RM：

$$P(w \mid M_R) = \sum_{i=1}^{k} P(D_i \mid Q) P(w \mid D_i)$$

$$P(D_i \mid Q) = \frac{P(Q \mid D)^{\alpha}}{\sum_{j=1}^{k} P(Q \mid D_j)^{\alpha}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (adjust\ \alpha)$$

- HMM+RM：

$$P^*(w \mid S_i) = (1-\lambda) P(w \mid S_i) + \lambda P(w \mid M_R) \ldots\ldots\ldots (adjust\ \lambda)$$

$$P(D \mid S_i) = \prod (1-\beta) P^*(w \mid S_i) + \beta P(w \mid C)$$

- wTMM

$$P(D|S_i) = \prod_{w \in D} \left( \sum_{l \in S_i} \alpha_l P(w|M_{w_l}) \right)^{c(w,D)}$$

$$= \prod_{w \in D} \left( \sum_{l \in S_i} \alpha_l \sum_{k=1}^{K} P(T_k|M_{w_l}) P(w|T_k) \right)^{c(w,D)}$$

$$\rightarrow \prod_{w \in D} \left( a \times \left( \sum_{l \in S_i} \alpha_l \sum_{k=1}^{K} P(T_k|M_{w_l}) P(w|T_k) \right) + (1-a) P(w|S_i) \right)^{c(w,D)}$$

# Data Corpus and Experiments(1/7)

- Data Corpus Ⅰ：
  - 蒐集自News 98新聞網 包含2001年8月1日至8月24日中午12:00~13:00的FM廣播新聞，相關統計資料如下：

| | |
|---|---|
| 新聞時間 | 2001年8月1日~2001年8月24日 |
| 新聞數 | 200則 |
| 總長度 | 1.61小時 |
| 平均每則新聞長度 | 28.96秒 |
| 總大小（人工轉寫） | 約31仟字 |
| 平均每則新聞大小（人工轉寫） | 約157字 |

  - 200則廣播新聞共分爲自動轉寫（Automatic Transcription）與人工轉寫（Human Transcription）兩種資料集
  - 辨識字正確率達85.83%
  - 自動摘要評估的標準答案爲三位國立台灣大學文學院大三以上的學生，分別對這200則廣播新聞的人工轉寫做摘要
    - 句排名摘要（Extraction）
    - 依特定比例重寫摘要（Abstraction）

# Data Corpus and Experiments(2/7)

- Data Corpus Ⅱ：
  - 使用MATBN新聞語料，由2001年十一月及2002年至八月份的新聞中選取其中205篇做爲自動摘要的測試語料，相關統計資料如下：

| | |
|---|---|
| 新聞時間 | 2001年11月~2002年8月 |
| 新聞數 | 205則 |
| 總長度 | 7.54小時 |
| 平均每則新聞長度 | 132.38秒 |
| 總大小（人工轉寫） | 約124仟字 |
| 平均每則新聞大小（人工轉寫） | 約604字 |

  - 自動摘要評估的標準答案由三位大三以上的政大新聞系同學分別對這205則新聞的人工轉寫文件作摘要，分爲句排名（Extraction）與依特定比例重寫（Abstraction）
  - 辨識正確率：

| 1.整篇(內外場) | 2.主播(內場) | 3.記者(外場) | 4.受訪者(外場) |
|---|---|---|---|
| 61.31 | 61.31 | 78.08 | 4.48 |

# Data Corpus and Experiments(3/7)

- Training Data：
  - 中央通訊社（Central News Agency）西元2001年08月且型態屬於故事（type=“story”）的文字新聞做爲訓練語料庫Ⅰ
  - 中央通訊社（Central News Agency）西元2001年11月~ 2002年08月且型態屬於故事（type=“story”）的文字新聞爲訓練語料庫Ⅱ
  - 每一篇新聞皆含有文件與標題兩部份，其內容包括國內外及大陸文教、交通、社會、財經、國會、影劇、醫藥衛生、體育及地方新聞

# Data Corpus and Experiments(4/7)

- Experiments
  - Data Corpus Ⅰ
    - 人工轉寫

| 摘要比例 | VSM | LSA | eLSA | DIM | HMM | TMM | Sig_Score | Random |
|---|---|---|---|---|---|---|---|---|
| 10% | 0.3591 | 0.3721 | 0.3647 | 0.3626 | 0.3875 | 0.3884 | 0.3584 | 0.1572 |
| 20% | 0.3898 | 0.4009 | 0.4111 | 0.3948 | 0.4383 | 0.4328 | 0.4182 | 0.1922 |
| 30% | 0.4572 | 0.4410 | 0.4868 | 0.4617 | 0.4827 | 0.4886 | 0.4963 | 0.2915 |
| 50% | 0.6217 | 0.5714 | 0.6365 | 0.6415 | 0.6551 | 0.6497 | 0.6540 | 0.4744 |
| 70% | 0.7540 | 0.7001 | 0.7605 | 0.7740 | 0.7801 | 0.7763 | 0.7931 | 0.6129 |

    - 自動轉寫

| 摘要比例 | VSM | LSA | eLSA | DIM | HMM | TMM | Sig_Score | Random |
|---|---|---|---|---|---|---|---|---|
| 10% | 0.2845 | 0.2755 | 0.2833 | 0.2498 | 0.2989 | 0.2994 | 0.2645 | 0.1122 |
| 20% | 0.3110 | 0.2911 | 0.3182 | 0.2917 | 0.3295 | 0.3314 | 0.3058 | 0.1263 |
| 30% | 0.3435 | 0.3081 | 0.3474 | 0.3378 | 0.3670 | 0.3671 | 0.3431 | 0.1834 |
| 50% | 0.4565 | 0.4070 | 0.4737 | 0.4666 | 0.4743 | 0.4753 | 0.4847 | 0.3096 |
| 70% | 0.5482 | 0.4930 | 0.5559 | 0.5575 | 0.5633 | 0.5627 | 0.5700 | 0.4252 |

# Data Corpus and Experiments(5/7)

- Experiments：ROUGE measure
  - Data Corpus Ⅱ
    - 人工轉寫

| 摘要比例 | VSM | LSA | eLSA | HMM | DIM | Random | Sig_Score | TMM_8 |
|---|---|---|---|---|---|---|---|---|
| 10% | 0.2838 | 0.2368 | 0.3582 | 0.3136 | 0.1744 | 0.0801 | 0.1332 | 0.3248 |
| 20% | 0.4218 | 0.3588 | 0.4788 | 0.4551 | 0.3484 | 0.2066 | 0.2594 | 0.4654 |
| 30% | 0.4934 | 0.4538 | 0.5365 | 0.5195 | 0.4962 | 0.2871 | 0.3360 | 0.5241 |
| 50% | 0.6583 | 0.6409 | 0.6989 | 0.6909 | 0.6958 | 0.4821 | 0.5503 | 0.6935 |
| 70% | 0.7916 | 0.7820 | 0.8112 | 0.8193 | 0.8127 | 0.6582 | 0.7359 | 0.8209 |

    - 自動轉寫

| 摘要比例 | VSM | LSA | eLSA | HMM | DIM | random | Sig_Score | TMM_8 |
|---|---|---|---|---|---|---|---|---|
| 10% | 0.2065 | 0.1567 | 0.2154 | 0.2142 | 0.1503 | 0.0793 | 0.0698 | 0.2173 |
| 20% | 0.2538 | 0.2174 | 0.2621 | 0.2678 | 0.2369 | 0.1347 | 0.1438 | 0.2722 |
| 30% | 0.2992 | 0.2677 | 0.3091 | 0.3108 | 0.2906 | 0.1724 | 0.2052 | 0.3123 |
| 50% | 0.3821 | 0.3418 | 0.3851 | 0.3827 | 0.3801 | 0.2623 | 0.3180 | 0.3839 |
| 70% | 0.4165 | 0.3919 | 0.4195 | 0.4176 | 0.4207 | 0.3354 | 0.3968 | 0.4209 |

# Data Corpus and Experiments(6/7)

- wTMM result (Model1)

Model1: Search all Data Corpus

Model2: Search Day Data

| TD_it100 | Mix_2 | Mix_4 | Mix_8 | Mix_16 | Mix_32 | Mix_64 |
|---|---|---|---|---|---|---|
| 0.3884 | 0.3905 | 0.4007 | 0.4097 | 0.4140 | 0.3980 | 0.4039 |
| 0.4328 | 0.4330 | 0.4283 | 0.4389 | 0.4354 | 0.4220 | 0.4339 |
| 0.4886 | 0.4919 | 0.4871 | 0.4959 | 0.4888 | 0.4836 | 0.4890 |
| 0.6497 | 0.6421 | 0.6409 | 0.6391 | 0.6500 | 0.6454 | 0.6446 |
| 0.7763 | 0.7715 | 0.7691 | 0.7720 | 0.7702 | 0.7682 | 0.7631 |
| **TD_it50** | Mix_2 | Mix_4 | Mix_8 | Mix_16 | Mix_32 | Mix_64 |
| | 0.3905 | 0.3956 | 0.4077 | 0.4143 | 0.4006 | 0.4091 |
| | 0.4331 | 0.4257 | 0.4389 | 0.4360 | 0.4170 | 0.4385 |
| | 0.4914 | 0.4901 | 0.4948 | 0.4920 | 0.4825 | 0.4989 |
| | 0.6421 | 0.6402 | 0.6417 | 0.6462 | 0.6373 | 0.6447 |
| | 0.7715 | 0.7647 | 0.7710 | 0.7706 | 0.7657 | 0.7625 |

| SD_it100 | Mix_2 | Mix_4 | Mix_8 | Mix_16 | Mix_32 | Mix_64 |
|---|---|---|---|---|---|---|
| 0.2994 | 0.3056 | 0.2944 | 0.3171 | 0.31107 | 0.3208 | 0.3162 |
| 0.3314 | 0.3323 | 0.3301 | 0.3436 | 0.33622 | 0.3345 | 0.3398 |
| 0.3671 | 0.3734 | 0.3689 | 0.3704 | 0.37892 | 0.3747 | 0.3709 |
| 0.4753 | 0.4735 | 0.4733 | 0.4752 | 0.4691 | 0.4721 | 0.4738 |
| 0.5627 | 0.5615 | 0.5634 | 0.5606 | 0.56274 | 0.5617 | 0.5582 |

# Data Corpus and Experiments(7/7)

- wTMM result (Model2)

| TD_it100 | Mix_2 | Mix_4 | Mix_8 | Mix_16 | Mix_32 | Mix_64 |
|---|---|---|---|---|---|---|
| | 0.40073 | 0.3975 | 0.4003 | 0.4027 | 0.38989 | 0.3836 |
| | 0.43323 | 0.4304 | 0.4313 | 0.4306 | 0.40838 | 0.4065 |
| | 0.48456 | 0.4880 | 0.4903 | 0.4914 | 0.48176 | 0.4884 |
| | 0.63747 | 0.6456 | 0.6426 | 0.6492 | 0.64395 | 0.6418 |
| | 0.77498 | 0.7686 | 0.7645 | 0.7708 | 0.75988 | 0.7604 |
| TD_it50 | Mix_2 | Mix_4 | Mix_8 | Mix_16 | Mix_32 | Mix_64 |
| | 0.4007 | 0.3946 | 0.4020 | 0.4025 | 0.3978 | 0.3786 |
| | 0.4332 | 0.4266 | 0.4326 | 0.4320 | 0.4162 | 0.4019 |
| | 0.4842 | 0.4874 | 0.4915 | 0.4915 | 0.4869 | 0.4856 |
| | 0.6372 | 0.6419 | 0.6427 | 0.6479 | 0.6449 | 0.6463 |
| | 0.7762 | 0.7672 | 0.7654 | 0.7715 | 0.7618 | 0.7626 |

# Conclusions

- To survey some summary methods and Evaluations
- Using some method or model to improve summary results